

On correlation between protein secondary structure, backbone bond angles, and side-chain orientations

Martin Lundgren^{1,*} and Antti J. Niemi^{2,1,†}

¹*Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*

²*Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083,*

Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France

We investigate the fine structure of the *sp*3 hybridized covalent bond geometry that governs the tetrahedral architecture around the central C_α carbon of a protein backbone, and for this we develop new visualization techniques to analyze high resolution X-ray structures in Protein Data Bank. We observe that there is a correlation between the deformations of the ideal tetrahedral symmetry and the local secondary structure of the protein. We propose a universal coarse grained energy function to describe the ensuing side-chain geometry in terms of the C_β carbon orientations. The energy function can model the side-chain geometry with a sub-atomic precision. As an example we construct the C_α - C_β structure of HP35 chicken villin headpiece. We obtain a configuration that deviates less than 0.4 Å in root-mean-square distance from the experimental X-ray structure.

PACS numbers: 87.15.Cc 05.45.Yv 36.20.Ey

I: INTRODUCTION

Protein structure validation is based on various well tested and broadly accepted stereochemical paradigms. Methods such as *MolProbit* [1] and *Procheck* [2] and many others help crystallographers to find and fix potential problems during fitting and refinement. Stereochemical assumptions are also instrumental to structure prediction packages such as *Rosetta* and *I-Tasser* [3]. Likewise, they form the foundation for parameter determination in force fields such as *Charmm* and *Amber* [4] that aim to describe protein dynamics at atomic scale.

One of the paradigms is the transfersability assumption. It states that stereochemical restraints are universal and independent of the environment. Among its consequences are that the covalent bond geometry around the backbone C_α should seize a very precise tetrahedral *sp*3 hybridized shape. For example, the backbone

$$\tau_{NC} \equiv (N - C_\alpha - C)$$

bond angle should oscillate around a computable average value that depends only on the covalent bonds between the C_α and the N, C, H and C_β atoms in the trans-peptide group. In particular, at least to the leading order its value should not depend on the character of the secondary structure environment. Standard molecular dynamics force fields explicitly assume this to be the case. These force fields are based on a harmonic approximation where the bond angles κ oscillate with energy [4]

$$E_{bond} = \sum_{bonds} \omega_\kappa (\kappa - \kappa_0)^2 \quad (1)$$

Here ω_κ and κ_0 are parameters that are in general amino acid dependent. But these parameters are presumed to be independent of the geometry of the surrounding secondary structure. Instead, they are supposed to predict the local secondary structure environment.

The enormous success that has been enjoyed by the validation methods and structure prediction programs in resolving close to 80.000 crystallographic protein structures that are presently in Protein Data Bank (PDB) [5] is a clear manifestation that the various paradigms are valid to a good precision. However, with the advent of third-generation synchrotron sources of X-rays, there is now a small but rapidly expanding number of protein structures that are resolved with an ultrahigh sub-Angström resolution. The present, third-generation X-ray synchrotron sources such as ESRF in Grenoble and PETRA at DESY in Hamburg can already produce photons with wavelengths as short as 10 pico-meters. Thus it is in principle possible to obtain three dimensional protein structures with a comparable resolution. The next-generation sources of high brilliance X-ray beams such as the European X-Ray Free Electron Laser at DESY, will push protein X-ray crystallography to its extreme. These future experimental facilities can reach both ultrahigh spatial and temporal resolutions, with a fully coherent peak brightness that is many orders of magnitude higher than what can be obtained with the present third-generation synchrotron sources. The on-going experimental revolution in combination with the ever expanding need of higher precision for example in the study of protein-protein interactions, enzyme catalysis and search of causes for protein misfolding related diseases, are good incentives for us to scrutinize the level of precision in some of the paradigm assumptions on protein backbone geometry. And, if need be, to try and develop new theoretical concepts that aim to describe proteins at a precision that matches the highest present and near future experimental standards, in revealing the finer structures of folded proteins.

In fact, *ab initio* quantum mechanical calculations [6] and empirical studies [7]-[9] of protein backbone geometry have already disclosed that the backbone bond angle

$\tau_{NC} \equiv (N-C_\alpha-C)$ about the C_α carbons might oscillate quite substantially. The range of variations can be as large as 8.8° [8]. This corresponds to a shift of ~ 0.6 Å in the relative positioning of two consecutive C_α carbons. A deviation of this size from the ideal value can be subjected to experimental scrutiny in X-ray experiments that reach sub-Ångström resolution. Indeed, on the basis of existing data the authors [7]-[9] have already reported that the deviations in the values of the τ_{NC} angle are systematic, and in particular that these deviations reflect the local secondary structure.

The τ_{NC} angles are primarily affected by the backbone. As such, their values relate directly to the two standard Ramachandran angles, that form the basis for structure validation. As a consequence, the literature [6]-[9] has until now mainly concentrated on the effects that potential deviations of τ_{NC} from ideality have on the backbone geometry. Here we extend this analysis to the side-chains: The fluctuations in the lengths of the covalent bonds in the C_α tetrahedron are no more than around 0.1 Å which is much less than the potential ~ 0.6 Å shift in the relative positioning of two consecutive C_α carbons, due to τ_{NC} fluctuations [7]-[9]. This proposes that any deviation of τ_{NC} from its ideal value inevitably propagates to the side-chain dependent $\tau_{N\beta} \equiv (N-C_\alpha-C_\beta)$ and $\tau_{C\beta} \equiv (C-C_\alpha-C_\beta)$ bond angles, and this should lead to observable effects in the angular positions of the side-chain C_β atoms.

In this article we first analyze PDB data to find whether there are experimental variations in the tetrahedral angles around the C_α . In particular, we extend the analysis of [7]-[9] to study correlations between the side-chain dependent angles $\tau_{N\beta}$ and $\tau_{C\beta}$ that determine the C_β orientations, and the local secondary structure of the backbone. Since the side-chain atom positions are not easily described in terms of the backbone Ramachandran angles, we start by developing new visualization tools. In line with [7]-[9] we observe that the local secondary structure has a systematic effect on the relative tetrahedral position of the C_β carbon. We then proceed to utilize our visualization tools to develop theoretical arguments. We propose a coarse-grained framework that computes how the observed direction of the C_β evolves along the backbone. In particular, we argue that the direction of the C_β can be computed from the soliton solution of a discrete nonlinear Schrödinger (DNLS) equation. The DNLS soliton already shares a remarkable history with protein research [10]. Both the DNLS equation and its soliton solution were first introduced by Davydov to describe the propagation of energy along α -helices [11]. He also proposed that since the propagation leads to a local deformation of the protein shape, a trapped soliton is a natural cause for the protein to fold. Here we first argue on general grounds that the DNLS soliton solution can be utilized to determine the secondary structure dependence in the relative direction of the C_β atoms along

the backbone. We then consider an explicit example to illustrate our general arguments. The example we consider is the 35-residue subdomain of the villin headpiece with PDB code 1YRF. It is a paradigm protein that has been studied widely in theoretical approaches to protein folding.

II: VISUALIZATION OF THE C_α TETRAHEDRON

We start by visual analysis of crystallographic protein data in PDB. The goal is to reveal any secondary structure dependence in the values of τ_{NC} , and in the adjacent

$$\tau_{N\beta} \equiv (N - C_\alpha - C_\beta)$$

and

$$\tau_{C\beta} \equiv (C - C_\alpha - C_\beta)$$

bond angles. In order to minimize any bias, we inspect several subsets of PDB. These include the canonical one that comprises all PDB configurations with resolution 2.0 Å or better, and its two subsets with resolution better than 1.5 Å, and better than 1.0 Å. We also inspect a subset of the 2.0 Å set that contains only those proteins that have less than 30% sequence similarity. Our conclusions are independent of the data set, and for illustrative purposes we here use the canonical 2.0 Å set. There are presently over 30.000 such entries in PDB.

The Ramachandran angles are defined in terms of the backbone amide planes. As such they are not the most convenient ones for describing the side-chain geometry. Since both $\tau_{N\beta}$ and $\tau_{C\beta}$ relate to the side-chain geometry, we prefer to follow [12] and describe the folded protein structure in terms of the geometrically determined backbone discrete Frenet frames (DFF). These frames govern the entire backbone neighborhood, including the side-chains. But their construction involves *only* the C_α coordinates \mathbf{r}_i where $i = 1, \dots, N$ label the residues. As such, these frames then give a manifestly N, C_β and C independent, purely geometric description of the tetrahedral *sp*3 neighborhood of the C_α atoms.

The backbone tangent vectors are

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|} \quad (2)$$

The unit binormal vectors are

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} - \mathbf{t}_i}{|\mathbf{t}_{i-1} - \mathbf{t}_i|} \quad (3)$$

The unit normal vectors

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i \quad (4)$$

The orthogonal triplets $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ determine the discrete Frenet frame at each of the positions \mathbf{r}_i of the C_α carbons. Note that if the tangent vectors and the distances

between the C_α are known, we can reconstruct the entire C_α backbone using

$$\mathbf{r}_k = \sum_{i=1}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \cdot \mathbf{t}_i \quad (5)$$

For the initial condition we can utilize Galilean invariance to take $\mathbf{r}_1 = 0$, and \mathbf{t}_1 to point into the direction of the positive z -axis. In particular, (5) does not involve the vectors \mathbf{n}_i and \mathbf{b}_i .

We introduce the backbone bond angles

$$\cos \kappa_{i+1} \equiv \cos \kappa_{i+1,i} = \mathbf{t}_{i+1} \cdot \mathbf{t}_i \quad (6)$$

and the backbone torsion angles

$$\cos \tau_{i+1} \equiv \cos \tau_{i+1,i} = \mathbf{b}_{i+1} \cdot \mathbf{b}_i \quad (7)$$

Note that both angles are manifestly independent of the N, C_β and C atoms that are covalently bonded to the C_α atoms, in this sense they provide an unbiased set of coordinates for describing the positions of these atoms. If these angles are known, we can use

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos \kappa \cos \tau & \cos \kappa \sin \tau & -\sin \kappa \\ -\sin \tau & \cos \tau & 0 \\ \sin \kappa \cos \tau & \sin \kappa \sin \tau & \cos \kappa \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \equiv \mathcal{R}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \quad (8)$$

to iteratively construct the Frenet frame at position $i+1$ from the frame at position i . Once we have all the frames, we can proceed to construct the entire backbone using (5).

The bond and torsion angles have a natural interpretation in terms of the canonical latitude and longitude angles of a two-sphere \mathbb{S}^2 . In the sequel we find it useful to extend the range of κ_i to $[-\pi, \pi] \bmod(2\pi)$. But we introduce no change in the range of $\tau_i \in [-\pi, \pi] \bmod(2\pi)$. We compensate for this two-fold covering of \mathbb{S}^2 by the following discrete \mathbb{Z}_2 symmetry

$$\begin{aligned} \kappa_k &\rightarrow -\kappa_k \quad \text{for all } k \geq i \\ \tau_i &\rightarrow \tau_i - \pi \end{aligned} \quad (9)$$

It inverts the directions of the vectors \mathbf{n}_i and \mathbf{b}_i but has no effect on the \mathbf{t}_i and consequently leaves the backbone intact. For details we refer to [12].

We use the discrete Frenet frames to display each atom in the way, how the atom is seen on the surface of a sphere that surrounds an imaginary observer who roller-coasts the backbone along the C_α atoms, so that the gaze direction is always fixed towards the next C_α and with local orientation determined by the DFF frames [12].

In Figure 1 we show the statistical angular distribution of the backbone N and C atoms, and in Figure 2 we show

the same for the side-chain C_β atoms in our PDB data set as seen by the Frenet frame observer who moves through all the proteins in our data set. The sphere is centered at the C_α , and its radius coincides with the length of the (approximatively constant) covalent bond. We take the vector \mathbf{t} that points towards the next C_α to be in the direction of the positive z -axis, towards the north-pole of the sphere. With \mathbf{n} in the direction of positive x -axis we have a right-handed Cartesian coordinate system. We introduce the canonical spherical coordinates (θ, φ) to describe the distributions. The angle $\theta \in [0, \pi]$ measures latitude from the positive z -axis, hence it describes the distribution of the bond angles κ_i . The angle $\varphi \in [-\pi, \pi]$ measures longitude in a counterclockwise direction from the x -axis *i.e.* from the direction of \mathbf{n} towards that of \mathbf{b} , with $\varphi = 0$ at the x -axis. Consequently it describes the distribution of the torsion angles τ_i .

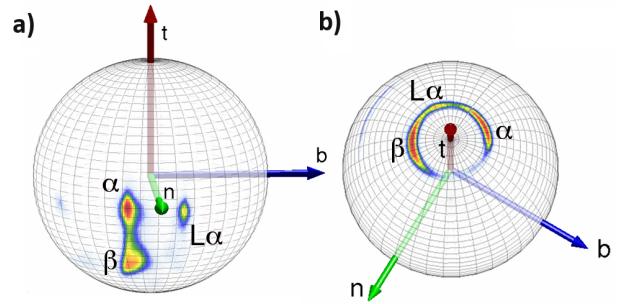


FIG. 1: (Color online) The directions of a) backbone N-atoms and b) backbone C-atoms as seen by a Frenet frame observer located at the C_α carbon which is at the center of the sphere. In a) the smaller, more point-like direction of backbone N atoms corresponds to the L- α Ramachandran region. The larger region forms a segment of the great circle $\varphi \approx -15^\circ$. Loops interpolate latitudinally between α -helices and β -sheets. In b) the directions of backbone C form a segment of a small circle around z -axis, with $\theta \approx 20^\circ$.

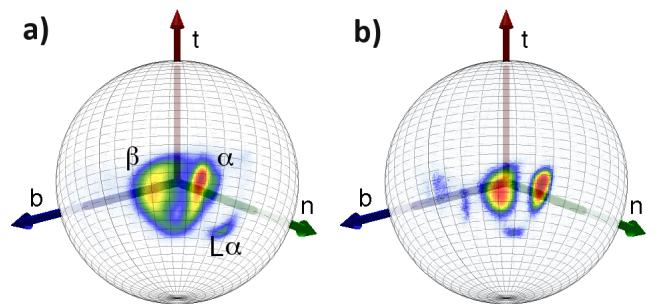


FIG. 2: (Color online) The distribution of the C_β directions in the Frenet frame. In a) we have all amino acids (including proline but excluding glycine that has no C_β). In b) we show only proline. Comparison between a) and b) exemplifies how the C_β direction can depend on the individual amino acid. We have chosen proline in b) as it is particularly interesting due to the way how it appears in Figure 3b).

We find that in the Frenet frame coordinate system, the N and C oscillations shown in Figures 1a) and 1b) are fully separated into the locally orthogonal θ and φ directions, respectively. This would certainly not be the case in a generic coordinate system. Furthermore, secondary structures such as α -helices, β -sheets, loops and left-handed α -regions are all clearly identifiable in Figures 1. Figure 2a) reveals how the N and C oscillations of Figure 1, through the covalent bonds that form the sp^3 tetrahedron around C_α , combine into a horseshoe (annulus) shaped nutation of C_β . As visible in the Figure, this nutation reflects the local secondary structure environment in an equally systematic manner as Figures 1.

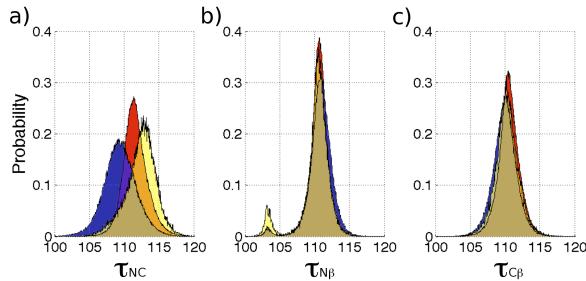


FIG. 3: (Color online) The normalized probability density angular distribution of the a) τ_{NC} , b) $\tau_{N\beta}$ and c) $\tau_{C\beta}$ angles in degrees, with α -helices in red (grey), β -strands in blue (dark grey), and 3/10 helices in yellow. The secondary peak in b) is due to *cis*-peptide prolines. Only in Figure a) are the different secondary structures visibly separated from each other. In Figures b) and c) they are practically fully overlapping, the only notable effect is the somewhat higher propensity of 3/10 helices in connection of *cis*-peptide prolines.

The pattern of angular separation in the N and C oscillations in Figure 1 reveals that the underlying ideal tetrahedral covalent symmetry around C_α is not transferable along the protein backbone [7]-[9].

We proceed to Figures 3 a)-c) where we plot the tetrahedral bond angles τ_{NC} , $\tau_{C\beta}$ and $\tau_{N\beta}$ jointly for the α -helices, β -strands and 3/10-helices. As in figure 1 the loops interpolate continuously between these regular secondary structures. The Figure 3a) clearly reveals that at the level of the τ_{NC} angle the transferability of the tetrahedral symmetry is absent in a systematic and secondary structure dependent manner. But neither $\tau_{C\beta}$ nor $\tau_{N\beta}$ show any sign whatsoever of secondary structure dependence. The distributions are practically the same, independently of the secondary structure. (The isolated small peak in Figure 3b) is due to proline *cis*-peptide groups.)

As such, the distribution in both Figures 3b) and 3c) is what we would expect in the case of ideal, transferable sp^3 tetrahedral symmetry. In each Figure the average value is around 111° and there are secondary structure independent fluctuations that are in line with quantum mechanical estimates and [6] and empirical studies [7]-

[9]. But the fact that τ_{NC} in Figure 3a) displays clearly visible and systematic secondary structure dependence makes it plain and clear that the paradigm of transferable tetrahedral covalent symmetry around the C_α carbon is absent. Furthermore, the way how transferability becomes violated reflects the wider secondary structure environment of the amino acid along the protein backbone.

Since the deviation from the ideal tetrahedral symmetry is organized in the same way how the proteins are folded, these two must share a common origin. But we do not have any physical explanation why the lack of ideal symmetry is only visible in τ_{NC} . We suspect this has to do with existing experimental refinement methods, the way how refinement tension is distributed between the backbone and side-chains. The high resolution crystallographic data which becomes available in future third and fourth generation experiments should help to clarify this.

III: SOLITONS AND SIDE-CHAINS

Any molecular dynamics approach to protein folding that we are familiar with, utilizes the harmonic approximation (1) for bond and torsion angles. Here κ_0 is in general amino acid dependent but secondary structure independent equilibrium value of the bond angle. But from Figure 3a) we observe, that for the τ_{NC} angle there are three different major equilibrium values. These equilibrium bond angle values are amino acid independent, but do depend in a nontrivial manner on the local secondary structure: The three peaks in Figure 3a) correspond to the α -helix, β -strand, and 3/10-helix while for a loop the values of the corresponding equilibrium κ_0 interpolates between these three ground state values. Since each of these secondary structures are characterized by a different equilibrium value of bond angle κ , to the leading order we may take κ_0 to be a function of κ . By expanding to leading order we get

$$\kappa_0 \rightarrow \kappa_0(\kappa) \approx \kappa_0^{(0)} + \kappa_0^{(1)} \cdot \kappa + \kappa_0^{(2)} \cdot \kappa^2 + \mathcal{O}(\kappa^3)$$

where the $\kappa^{(i)}$ are independent of the local value of κ . The first two terms simply renormalize the values of ω_κ and κ_0 in (1). But the third term is conceptually different. It introduces an anharmonic correction. We conclude that after redefinitions of the parameters, in the leading order the potential obtains the functional form

$$E_{bond} \sim q \cdot (\kappa^2 - m^2)^2 \quad (10)$$

We argue that based on the Figure 3a), the anharmonic corrections are already visible in the existing high resolution X-ray data. In order to answer the theoretical challenge that this poses, we propose to improve existing MD force fields, to account for the anharmonic corrections in the bond angle contribution.

A: Backbone energy

The presence of an anharmonic correction in the bond angle energy has important implications to the way how proteins fold. For this we start with a simple example. We consider the anharmonic potential (10) in the presence of a single coordinate x on a line. The Newton's equation is

$$m\ddot{x} = -\frac{dV}{dx}$$

We take the potential to have the form

$$V(x) = \frac{\sigma}{4} (x + b)^2 \cdot (x - a)^2 \quad (11)$$

We introduce

$$c = -\frac{1}{2}(b + a)$$

and define

$$y = x - \frac{1}{2}(a - b)$$

to arrive at the equation of motion

$$\ddot{y} = -\frac{\sigma}{m}y(y^2 - c^2)$$

which is essentially the continuum nonlinear Schrödinger equation (NLSE) [10], [11]. Note that the potential has the symmetric form (10). This equation is solved by

$$\begin{aligned} y(t) &= c \cdot \tanh\left[c\sqrt{\frac{\sigma}{2m}}(t - t_0)\right] \\ \Rightarrow x(t) &= y(t) + \frac{1}{2}(a - b) \\ &= -\frac{b \cdot e^{c\sqrt{\frac{\sigma}{2m}}(t-t_0)} - a \cdot e^{-c\sqrt{\frac{\sigma}{2m}}(t-t_0)}}{\cosh[c\sqrt{\frac{\sigma}{2m}}(t - t_0)]} \end{aligned} \quad (12)$$

This is the hallmark NLSE soliton configuration, so called dark soliton solution of the NLSE equation. It interpolates between the two uniform ground states at $x = a$ and $x = -b$ when $t \rightarrow \pm\infty$. The parameters a , b , t_0 and the combination

$$c\sqrt{\frac{\sigma}{2m}}$$

are the canonical ones that characterize the asymptotic values of $x(t)$ *i.e.* minima of the potential, and the size and location of the soliton.

For finite t the soliton (12) describes a configuration with an energy above the uniform ground state $x \equiv a$ (or $x \equiv b$). Nevertheless, it can not decay into $x \equiv a$

(or $x \equiv b$) through any kind of continuous finite energy transformation. In particular, a soliton configuration such as (12) can not be obtained from any approach that only accounts for perturbations that describe small localized fluctuations around the uniform background ground state.

In [13]-[15], it has been shown that the soliton profile (12) can be used to describe loops in folded proteins. For this one merges general geometric arguments with the concept of universality [16]-[19], to arrive at the following simplified, coarse-grained energy function for the backbone bond and torsion angles [20], [21],

$$\begin{aligned} E = & -\sum_{i=1}^{N-1} 2\kappa_{i+1}\kappa_i + \sum_{i=1}^N \left\{ 2\kappa_i^2 + q \cdot (\kappa_i^2 - m^2)^2 \right. \\ & \left. + \frac{d_\tau}{2} \kappa_i^2 \tau_i^2 - b_\tau \kappa_i^2 \tau_i - a_\tau \tau_i + \frac{c_\tau}{2} \tau_i^2 \right\} \end{aligned} \quad (13)$$

where κ_i and τ_i are the backbone bond and torsion angles (6), (7). Unlike force fields in molecular dynamics, the energy function (13) does not purport to explain the fine details of the atomic level mechanisms that give rise to protein folding. Instead, in line with general principles of effective Landau-Lifschitz theories it describes the properties of a folded protein backbone in terms of universal physical arguments. Indeed, according to the concept of universality [16]-[19] the energy function (13) can be viewed as the universal long distance limit that emerges from any atomic level energy function when the internal energy is coarse-grained to include only the backbone bond and torsion angles.

In order to construct the soliton solution, we start by introducing the τ -equation of motion

$$\frac{\partial E}{\partial \tau_i} = d_\tau \kappa_i^2 \tau_i - b_\tau \kappa_i^2 - a_\tau + c_\tau \tau_i = 0$$

$$\Rightarrow \tau_i[\kappa] = \frac{a_\tau + b_\tau \kappa_i^2}{c_\tau + d_\tau \kappa_i^2} \quad (14)$$

Notice that even though there are four parameters in (14) one of them, the overall scale, drops out. We then use (14) to eliminate the torsion angles, so that the energy for the bond angles becomes

$$E[\kappa] = -\sum_{i=1}^{N-1} 2\kappa_{i+1}\kappa_i + \sum_{i=1}^N 2\kappa_i^2 + V[\kappa_i] \quad (15)$$

where

$$\begin{aligned} V[\kappa] = & -\left(\frac{b_\tau c_\tau - a_\tau d_\tau}{d_\tau}\right) \cdot \frac{1}{c_\tau + d_\tau \kappa^2} \\ & - \left(\frac{b_\tau^2 + 8qm^2}{2b_\tau}\right) \cdot \kappa^2 + q \cdot \kappa^4 \end{aligned} \quad (16)$$

Because the first term contains κ in the denominator, its variation with κ is not that pronounced as the variation of the other two terms, which are proportional to the second and the fourth power of κ , respectively. Moreover, because $|\kappa| > 1$ radian for proteins, it turns out that the first term is small in value compared to the other terms. The second and third terms have then the functional form of the double well potential (10).

In applications to folded proteins the parameters values are such, that in the energy ground state both κ and τ acquire a non-vanishing value. In particular, since the functional form of (16) is similar to (10), (11) we can expect that there are soliton solutions:

Geometrically, a uniform constant value of the bond and torsion angles describes regular protein secondary structures. For example, the standard α -helix is

$$\alpha - \text{helix} : \begin{cases} \kappa \approx \frac{\pi}{2} \\ \tau \approx 1 \end{cases} \quad (17)$$

and for the standard β -strand we have

$$\beta - \text{strand} : \begin{cases} \kappa \approx 1 \\ \tau \approx \pi \end{cases} \quad (18)$$

The additional regular secondary structures including 3/10 helices, left-handed helices *etc.* are described similarly.

But in addition of constant value configurations, as in (11) there are also soliton solutions. In particular, since protein loops are structures that interpolate between different constant values such as (17), (18), this means that loops correspond to these soliton solutions [13]-[15]. In order to construct the relevant soliton, we introduce the generalized discrete nonlinear Schrödinger (DNLS) equation that derives from the energy (13). Variation of this energy *w.r.t.* κ_i and substitution of (14) gives

$$\kappa_{i+1} = 2\kappa_i - \kappa_{i-1} + \frac{dV[\kappa]}{d\kappa_i^2} \kappa_i \quad (i = 1, \dots, N) \quad (19)$$

where $\kappa_0 = \kappa_{N+1} = 0$. The exact soliton solution to the present discrete nonlinear Schrödinger equation is not known in a closed form. But numerical approximations can be easily constructed using the procedure described in [14]. Furthermore, whenever the first term in (16) is small as it is in the case of proteins, an excellent approximation [15] is obtained from the *naive* discretization of the continuum soliton (12),

$$\kappa_i = \frac{\mu_1 \cdot e^{\sigma_1(i-s)} - \mu_2 \cdot e^{-\sigma_2(i-s)}}{e^{\sigma_1(i-s)} + e^{-\sigma_2(i-s)}} \quad (20)$$

Here s is a parameter that determines the backbone site location of the center of the fundamental loop that is described by the soliton. The $\mu_{1,2} \in [0, \pi]$ are parameters, their values are entirely determined by the adjacent helices and strands: Away from the soliton center we have

$$\kappa_i \rightarrow \begin{cases} \mu_1 & i > s \\ -\mu_2 & i < s \end{cases}$$

and for α -helices and β -strands the $\mu_{1,2}$ values are determined by (17), (18). We remind that negative values of κ_i are related to the positive values by (9). Note that for $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ we recover the hyperbolic tangent. In this case the two regular secondary structures before and after the loop are the same. Moreover, *only* the (positive) σ_1 and σ_2 are intrinsically loop specific parameters, they specify the length of the loop and as in the case of the $\mu_{1,2}$, they are combinations of the parameters in (13).

Similarly, in the case of the torsion angle there is only one loop specific parameter in (14): The overall, common scale of the four parameters is irrelevant in (14), and two of the remaining three parameters characterize the regular secondary structures that are adjacent to the loop, as in (17), (18).

Entire protein loops can be constructed by combining together solitons (19), (14). In [22] it has been shown using the Ansatz (20) that over 92% of crystallographic PDB configurations can be constructed in terms of 200 explicit soliton profiles. The solitons of the DNLS equation can thus be interpreted as the modular building blocks of folded proteins.

B: Side-chain energy

We proceed to extend the energy function (13) so that it models the deviations from the paradigm tetrahedral symmetry around the C_α atoms: The Figures 1a) and 1b) reveal that the directions of the backbone N and C atoms oscillate in the latitudinal (θ) and longitudinal (φ) directions respectively, on the surface of the sphere that surrounds the corresponding C_α atom. The covalent bond structure that forms the sp^3 tetrahedron of the C_α atom combines these two oscillations into the annulus (horseshoe) shaped C_β nutation of Figure 2a). Consequently the natural dynamical variable that describes the nutation of C_β on the surface of the sphere is the canonically parametrized three component unit vector

$$\mathbf{u} = \begin{pmatrix} \sin \theta \cdot \cos \varphi \\ \sin \theta \cdot \sin \varphi \\ \cos \theta \end{pmatrix} \quad (21)$$

In order to account for the C_β nutation contribution to the protein free energy, we then need to augment (13) by terms that engage the additional variables (θ_i, φ_i).

The latitude angle θ_i is counted from the direction of the corresponding Frenet frame tangent vector \mathbf{t}_i . Consequently it remains invariant under the rotations of the local Frenet frames around the direction of \mathbf{t}_i [12]. Thus it can only couple to other frame rotation invariant quantities. There are two natural terms,

$$|\mathbf{t}_i \times \mathbf{u}_i| = \sin \theta_i$$

and

$$\mathbf{t}_i \cdot \mathbf{u}_i = \cos \theta_i$$

In the leading order we only account for local interactions. When we also demand invariance under the \mathbb{Z}_2 gauge transformation (9) we conclude that to the leading order the corresponding free energy contribution should have the form

$$E_\theta = \sum_{i=1}^N f_i(\kappa_i^2) |\mathbf{t}_i \times \mathbf{u}_i| + g_i(\kappa_i^2) \mathbf{t}_i \cdot \mathbf{u}_i + \dots \quad (22)$$

According to Figure 2a) the range of variations in θ_i are quite small and we estimate that the center of the annulus-like region is near

$$\langle \theta \rangle \approx 113.4^\circ$$

We Taylor expand (21) around this value so that we have to the leading order

$$E_\theta = \sum_{i=1}^N \left\{ \frac{d_\theta}{2} \kappa_i^2 \theta_i^2 - b_\theta \kappa_i^2 \theta_i - a_\theta \theta_i + \frac{c_\theta}{2} \theta_i^2 \right\} + \dots \quad (23)$$

The ensuing equation of motion is

$$\theta_i = \frac{a_\theta + b_\theta \kappa_i^2}{c_\theta + d_\theta \kappa_i^2} \quad (24)$$

As in the case of (14) we conclude that the overall scale of the parameters drops out and this leaves us with three independent parameters. In the case of a short loop that we can model in terms of a single soliton like (20), two of the parameters become determined by the value of θ_i in the regular secondary structures that are adjacent to the loop. This leaves us with only one loop specific parameter.

The longitude φ_i in (21) is measured from the direction of the Frenet frame normal vector \mathbf{n}_i . Under the local rotations of the Frenet frames

$$\begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \end{pmatrix} \rightarrow \begin{pmatrix} \cos \Delta_i & \sin \Delta_i \\ -\sin \Delta_i & \cos \Delta_i \end{pmatrix} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \end{pmatrix}$$

around the tangent vectors \mathbf{t}_i by an angle Δ_i we then have

$$\varphi_i \rightarrow \varphi_i + \Delta_i$$

Thus we may couple φ to the torsion angle as follows,

$$\varphi_i + \sum_{k=1}^i \tau_k$$

This combination is invariant under the local rotations of the Frenet frame around \mathbf{t}_i . Since the τ_k depend on the backbone angles according to (14), we can again Taylor expand the ensuing energy contribution. From Figure 2a) we estimate that for the center of the annulus

$$\langle \varphi \rangle \approx 139.5^\circ$$

Following (23) we then Taylor expand the φ contribution to free energy around this value to conclude that to the leading order we have (in Frenet frames)

$$E_\varphi = \sum_{i=1}^N \left\{ \frac{d_\varphi}{2} \kappa_i^2 \varphi_i^2 - b_\varphi \kappa_i^2 \varphi_i - a_\varphi \varphi_i + \frac{c_\varphi}{2} \varphi_i^2 \right\} + \dots \quad (25)$$

The equation of motion has the same functional form with (14), (24)

$$\varphi_i = \frac{a_\varphi + b_\varphi \kappa_i^2}{c_\varphi + d_\varphi \kappa_i^2} \quad (26)$$

Again only three of the four parameters in φ are independent, the overall scale drops out.

We confirm that the functional forms (23) and (25) are in line with the annulus-like (horseshoe-like) form of the C_β nutation in Figure 2a). For this we stereographically project the C_β distribution in Figure 2a) onto the complex plane. Despite the nonlinear nature of the standard stereographic projection the annulus-like shape is more or less retained. Let r be the approximate radius of a thin annulus on the complex plane and let (θ_0, φ_0) be the location of its center. In the limit where the corrections to the round circular profile of the thin annulus become small we can determine the approximate form of the C_β nutation region from

$$(\tan \theta e^{i\varphi} - \tan \theta_0 e^{i\varphi_0})(\tan \theta e^{-i\varphi} - \tan \theta_0 e^{-i\varphi_0}) = r^2 \quad (27)$$

Note that this is invariant under local frame rotations. We re-write θ in (24) as follows,

$$\theta = \theta_0 + \frac{1}{c + d\kappa^2}$$

We substitute this into (27) and Taylor expand to find that to *leading order* it makes sense to parametrize φ by an expression of the functional form (26).

We make the following remark: When we combine (13), (23) and (25) we arrive at the total energy

$$E = - \sum_{i=1}^{N-1} 2 \kappa_{i+1} \kappa_i + \sum_{i=1}^N \left\{ 2 \kappa_i^2 + q \cdot (\kappa_i^2 - m^2)^2 \right\} \quad (28)$$

$$+ \sum_{i=1}^N \left\{ \frac{d_\tau}{2} \kappa_i^2 \tau_i^2 - b_\tau \kappa_i^2 \tau_i - a_\tau \tau_i + \frac{c_\tau}{2} \tau_i^2 \right\} \quad (29)$$

$$+ \sum_{i=1}^N \left\{ \frac{d_\theta}{2} \kappa_i^2 \theta_i^2 - b_\theta \kappa_i^2 \theta_i - a_\theta \theta_i + \frac{c_\theta}{2} \theta_i^2 \right\} + \dots \quad (30)$$

$$+ \sum_{i=1}^N \left\{ \frac{d_\varphi}{2} \kappa_i^2 \varphi_i^2 - b_\varphi \kappa_i^2 \varphi_i - a_\varphi \varphi_i + \frac{c_\varphi}{2} \varphi_i^2 \right\} \quad (31)$$

We have already established that protein backbones can be described in terms of soliton solutions to (28), (29). According to (14), (24), (26) the presence of (30) and (31) does not change the functional form of the effective κ_i energy (15), (16), all three variables $(\tau_i, \theta_i, \varphi_i)$ are similarly slaved to the bond angles κ_i . In particular, from Figure 2a) we conclude that the contribution of (30) and (31) to the full energy must be minuscule: The range of variations in the variables (θ_i, φ_i) is relatively small. (This is *not* the case with τ_i , see for example Figure 4 below.) Thus the values of (30) and (31) show very little variation, and in comparison to (29) these two terms can be treated as if they were tiny perturbations.

Indeed, in the case of proteins the two terms (30) and (31) make no contribution to the total energy that we are able to observe. The variables (θ_i, φ_i) are entirely slaved by the DNLS soliton profile of the backbone bond angles κ_i . Since the direction of the vector (21) that specifies the position of the C_β carbon is slaved to κ_i , the deviation from the ideal tetrahedral symmetry in the C_α covalent bond geometry is determined by the local secondary structure environment of the amino acid.

C: Comment on parameters

The energy function (28)-(31) introduces eleven essential parameters, when we account for the overall scales in (14), (24), (26). According to [22], no more than 200 different parameter sets are needed to describe over 92% of high resolution structures in PDB with a precision of around 0.6 Å in RMSD for the C_α . The solitons are like modular components from which the folded proteins are built. At the moment we do not have a method to compute the parameters directly from the sequence. However, even in its present form the approach can be subjected to a stringent experimental scrutiny: A typical super-secondary structure described by a soliton such as a helix-loop-helix consists of around 15 amino acids. If we assume that the bond lengths are fixed, this leaves us with 60 unknown coordinates for the C_α and C_β atoms. Since there are only 11 essential parameters in (28)-(31), we have a highly under-determined set of equations. Consequently the model is predictive, a comparison with experimental structures is directly testing the physical principles on which (28)-(31) is based, *even though* we are not yet able to compute the parameters from the sequence.

IV: EXAMPLE: VILLIN HEADPIECE HP35

As an example we consider the chicken villin headpiece subdomain HP35. We use the x-ray structure with PDB code 1YRF. The HP35 is a naturally existing 35-residue protein with three α -helices separated from each other by two loops. It continues to be the subject of very ex-

tensive studies both experimentally [23]-[25] and *in silico* [26]-[29], and [29] reports on a molecular dynamics construction with overall backbone RMSD accuracy around one Ångström.

In Figure 4 we have the (κ_i, τ_i) spectrum that we compute from the PDB data of 1YRF. In the Figure 4a) we use the standard convention that bond angles take values in the range $[0, \pi]$. In the Figure 4b) we have extended the range to $[-\pi, \pi]$. This introduced the \mathbb{Z}_2 gauge transformation structure (9). In Figure 4b) we have applied the gauge transformation to disclose the solitons. We clearly have two solitons with the DNLS profile (20), separated from each other by regions with $\kappa \approx \pm 1.57$ and $\tau \approx 1$ corresponding to the α -helix (17). Notice the irreg-

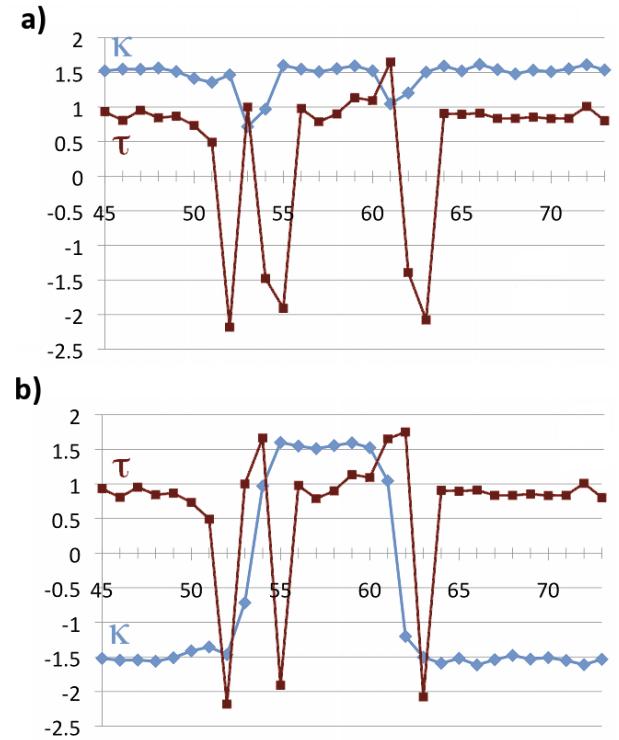


FIG. 4: (Color online) The profile of κ_i (light blue) and τ_i (dark red) along the 1YRF backbone. We use PDB indexing of the sites. In a) the κ_i are restricted to $[0, \pi]$ and in b) this region is extended to $[-\pi, \pi]$ using (9).

ular structure of the torsion angle τ_i in the loop (soliton) regions. *A priori* we expect from (14) that the torsion angle should have a regular profile. However, the numerical values that we compute from (14) are not restricted to the fundamental range $\tau_i \in [-\pi, \pi]$, they can take values beyond this range. The irregular structure of τ_i follows when we convert the values to the fundamental range, using 2π periodicity of τ_i in the discrete Frenet equation (8). Similarly we observe slight irregularity in the κ_i profile. This can also be removed if we allow κ_i to take values beyond $[-\pi, \pi]$ and use the 2π periodicity. But in the case of 1YRF the improvement in the precision turns

out to be very small, and consequently we search for a solution of (19) by assuming that $\kappa_i \in [-\pi, \pi]$.

In Figure 5 we show the distribution of the side-chain angles (θ_i, φ_i) in YRF, by plotting the tips of the unit vector (21) on the two-sphere of Figure 2. As expected, they are located in the α -helix region of Figure 2a) except along the loops, where they are located outside of the regular structure regions.

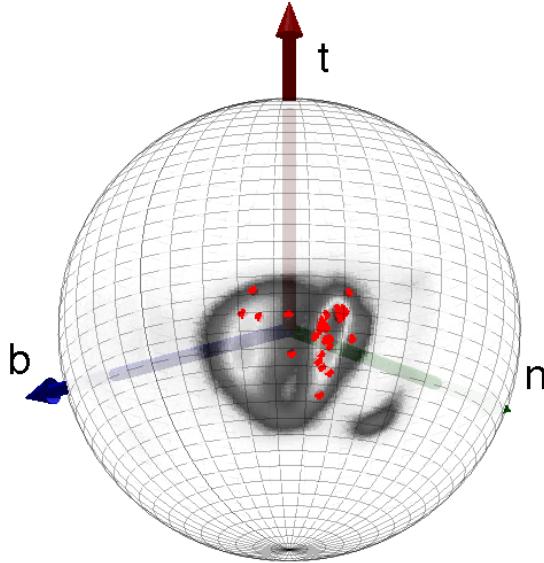


FIG. 5: (Color online) The directional distribution of the side-chain angles (θ_i, φ_i) . The background coincides with the annulus in Figure 2a).

We start by solving the classical equations of motion for κ_i from (19). We then construct the remaining variables $(\tau_i, \theta_i, \varphi_i)$ in terms of κ_i using (14), (24) and (26); Since the (θ_i, φ_i) contributions to the κ_i potential (16) are minuscule, we ignore the corresponding parameters in constructing the solution to the DNLS equation for κ_i . We use the iterative algorithm and procedure described in [30], [14], and our results are summarized in Figure 6 and Table 1. We have been able to substantially improve the accuracy reported in [14], in particular for the first soliton. We now reach a RMSD accuracy less than 0.4 Å even when we include the side-chain C_β atoms. The result is clearly within the Debye-Waller fluctuation distance regime that we compute from the experimental B-factors in the PDB data.

In Figure 6a) we display the distance between the computed and the experimentally measured C_α atoms (excluding the N and C terminals). The shaded region in Figure 6a) describes the 0.15 Å zero point fluctuations [22] around our solitons. For comparison, we also display the experimental Debye-Waller B-factor fluctuation distances, obtained from the PDB data. Except for the end point of soliton 1 (residue 58), our soliton solutions

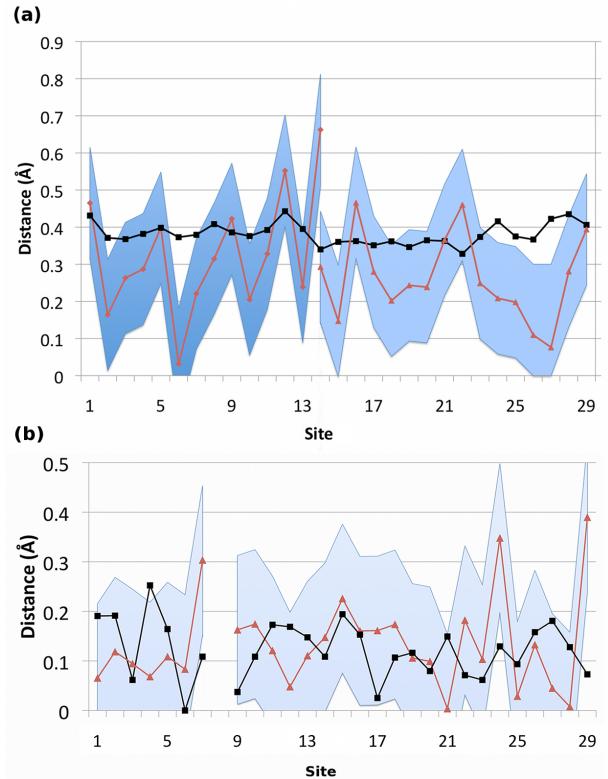


FIG. 6: (Color online) Comparison between our soliton solutions in red (gray) and the experimental B-factor fluctuation distance of PDB data for 1YRF (black) along the backbone. In Figure a) for the C_α , and in Figure b) for the C_β where the experimental accuracy is estimated from (32). The shaded region describes the 0.15 Å zero point fluctuations around solitons. The cut in Figure a) at sites 13-14 is where the two solitons overlap (Phe-58 in PDB), and the empty space in Figure b) is due to glycine that has no C_β .

describe the backbone well within the limits of experimental accuracy.

In Figure 6b) we present our results for the C_β nutation, in comparison with the experimental data. We also present an estimate for the experimental uncertainties that we estimate as follows: The experimental B-factors give an estimate for the *absolute* fluctuation distance around the measured position. But now we are interested in estimating the (much smaller) *relative* error in the position of C_β with respect to the position of the ensuing C_α . For this we introduce the *relative* B-factor

$$B_{rel} = |B_\alpha - B_\beta| \quad (32)$$

In Figure 6b) we display the ensuing fluctuation distances that we have computed from the Debye-Waller relation using (32) in lieu of the B-factor. The precision of our computed results compare well with these experimental relative B-factor errors: For most of the sites the difference is no more than the 0.15 Å estimate for zero point fluctuations.

parameter	soliton-1	soliton-2
q_1	0.459712	0.995867
q_2	4.5533320	9.408796
m_1	1.504535	1.550322
m_2	1.512836	1.535081
a_τ	9.5752137e-9	7.840467e-6
b_τ	-676965e-11	-4.973244e-9
c_τ	4.875744e-9	4.2733696e-6
d_τ	-2.917129e-9	-2.431388e-6
a_θ	1.514770	1.322495
b_θ	-0.0017952	-0.018619
d_θ	0.0420877	6.930946e-8
a_φ	0.544859	0.3594184
b_φ	5.66111e-5	3.83253e-4
d_φ	-0.1845828	-0.226012
RMSD (Å)	0.38	0.32

TABLE I: Parameter values for the two-soliton solution that describes the two loops of 1YRF with a combined 0.39 Å accuracy for both C_α and C_β atoms. The displayed RMSD values are for the individual solitons. The soliton-1 is located at Glu-45 - Phe-58 and the soliton-2 is located at Phe-58 - Lys-73. We utilize scale invariance to set all $c_\theta = c_\varphi = 1$. The result has sensitivity to the accuracy of parameters, because a folded protein is a piecewise linear polygonal chain with a positive Liapunov exponent.

Finally, Figure 7 shows our soliton solution together with the 1YRF configuration in PDB.

SUMMARY

In conclusion, the paradigm assumption that the tetrahedral covalent symmetry around the backbone C_α carbons is transferable, is correct to a good precision. However, with the advent of third and fourth generation X-ray sources there is now a rapid growth in the number of protein structures with sub-Ångström resolution. This makes it possible to scrutinize small corrections to this paradigm. We have found, that the backbone $N-C_\alpha-C$ bond angle shows systematic deviations from the ideal value, in a manner that is in direct correspondence with the corresponding secondary structure environment. We have investigated how this effect propagates to the orientation of the C_β carbon. We have found that the angular orientations of the C_β carbon similarly deviate from their ideal values, in a manner which is in a one-to-one correspondence with the underlying secondary structure environment.

We have presented a simple energy function that is based on the concept of universality, to model the secondary structure dependence in the C_β orientations. As an example, we have constructed the $C_\alpha-C_\beta$ backbone of HP35 villin, where we reach an accuracy that matches the experimental B-factor fluctuation distances. We propose that our observations and theoretical proposals could

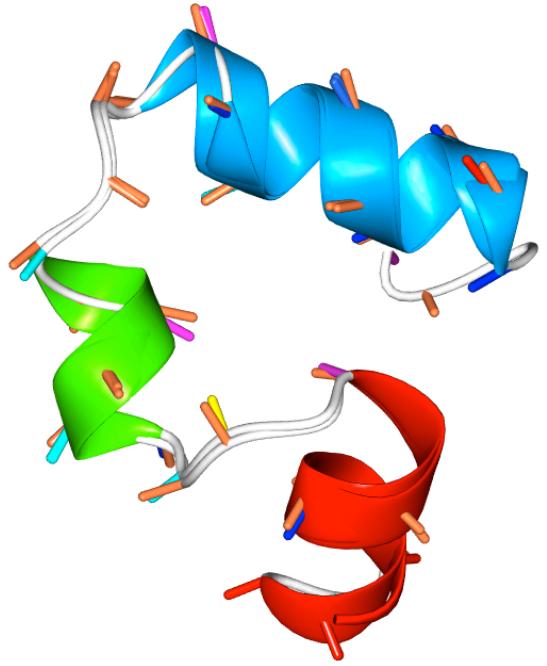


FIG. 7: (Color online) A cartoon comparison of HP35 with our soliton solution summarized in Table 1. The combined C_α and C_β root-mean-square distance is 0.39 Å which equals the experimental Debye-Waller B-factor fluctuation distance for the central carbons.

form a basis for the development of both more accurate refinement tools for experimental data analysis, and of more precise theoretical and computational MD force fields, to model the atomic level structure and dynamics of folded proteins.

* Electronic address: Martin.Lundgren@physics.uu.se

† Electronic address: Antti.Niemi@physics.uu.se

- [1] I.W. Davis *et.al.*, *Nucl. Acids Res.* **35** 235 (2007)
- [2] R.A. Laskowski *et.al.*, *J. Biomol. NMR* **8**, 477 (1996)
- [3] X. Qu, R. Swanson, R. Day, J. Tsai, *Curr. Protein Pept. Sci.* **10**, 270 (2009)
- [4] P.L. Freddolino, C.B. Harrison, Y. Liu, Y. Schulten, *Nature Phys.* **6**, 751 (2010)
- [5] H.M. Berman *et.al.*, *Nucl. Acids Res.* **28**, W375 (2007)
- [6] L. Schäfer, M. Cao, *Journ. Mol. Struc.* **333**, 201 (1995)
- [7] P.A. Karplus, *Prot. Sci.* **5**, 1406 (1996)
- [8] D.S. Berkholz, M.V. Shapovalov, R.L. Dunbrack Jr., P. A. Karplus, *Structure* **17**, 1316 (2009)
- [9] W.G. Touw, G. Vriend, *Acta Cryst.* **D66**, 1341 (2010)
- [10] P.G. Kevrekidis, *The Discrete Nonlinear Schrödinger Equation: Mathematical Analysis, Numerical Computations and Physical Perspectives* (Springer-Verlag, Berlin, 2009)
- [11] A.S. Davydov, *Journ. Theor. Biol.* **66**, 379 (1977)

- [12] S. Hu, M. Lundgren, A.J. Niemi, Phys. Rev. **E83**, 061908 (2011)
- [13] M. Chernodub, S. Hu, A.J. Niemi, Phys. Rev. **E82**, 011916 (2010)
- [14] N. Molkenthin, S. Hu, A.J. Niemi, Phys. Rev. Lett. **106**, 078102 (2011)
- [15] S. Hu, A. Krokhotin, A.J. Niemi, X. Peng, Phys. Rev. **E83**, 041907 (2011)
- [16] B. Widom, J. Chem. Phys. 43, 3892 (1965).
- [17] L.P. Kadanoff, Physics 2, 263 (1966).
- [18] K.G. Wilson, Phys. Rev. B4, 3174 (1971).
- [19] M.E. Fisher, Rev. Mod. Phys. 46, 597 (1974).
- [20] A.J. Niemi, Phys. Rev. **D67**, 106004 (2003)
- [21] U.H. Danielsson, M. Lundgren, A.J. Niemi, Phys. Rev. **E82**, 021910 (2010)
- [22] A. Krokhotin, A.J. Niemi, X. Peng, Phys. Rev. **E85**, 031906 (2012)
- [23] J. Meng, D. Vardar, Y. Wang, H.C. Guo, J.F. Head, C.J. McKnight, *Biochemistry* **44**, 11963 (2005)
- [24] T.K. Chiu, J. Kubelka, R. Herbst-Irmer, W.A. Eaton, J. Hofrichter, D.R. Davies, *Proc. Natl. Acad. Sci. U.S.A* **102**, 7517 (2005)
- [25] L. Wickstrom, Y. Bi, V. Hornak, D.P. Raleigh, C. Simmerling, *Biochemistry* **46**, 3624 (2007)
- [26] D.L. Ensign, P.M. Kasson, V.S. Pande, *J. Mol. Biol.* **374**, 806 (2007)
- [27] H. Lei, Y. Duan, *J. Mol. Biol.* **370**, 196 (2007)
- [28] P.L. Freddolino, K. Schulten, *Biophys. Journ.* **97**, 2338 (2009)
- [29] D.E. Shaw *et.al.*, *Science* **330**, 341 (2010)
- [30] M. Herrmann, *Applic. Anal.* **89**, 1591 (2010)